视频检索技术—从内容到上下文

曹娟 张勇东 李锦涛

摘要：随着视频采集设备的普及以及 Web2.0 技术的出现，互联网上的视频数据迅猛增长。如何从大规模视频数据中检索到用户需要的视频，正是视频检索技术所要解决的问题。本文对视频检索技术进行了综述，主要对基于内容的视频拷贝检测技术，基于概念的语义视频检索技术，以及基于上下文信息的网络视频分析技术进行了介绍。同时，本文也简要介绍了本课题组在视频拷贝检测，语义视频检测，以及网络视频分析方面的研究进展。

关键词：视觉特征 上下文信息 拷贝检测 视频检索 视频话题发现和推荐

1 引言

随着多媒体技术和网络技术的发展，视频已经成为人们日常生活中发布信息和获取信息的主要载体之一。2009 年 9 月，著名视频分享网站 YouTube 每分钟大约有 20 小时的新视频数据上传；根据中国互联网络信息中心报告，2010 年中国网络视频用户规模达到 2.84 亿人，占网民总数的 62.1%。面对网络视频及其用户的爆炸式增长，迫切需要高效的视频检索技术，帮助人们在大规模网络视频数据中快速、准确地找到所需要的视频内容。

视频检索技术在不同层次具有不同的表现形式。如下图所示，针对技术人员来说，视频检索技术包含视频结构化、特征提取、高维索引、 相似度计算、检索结果排序等核心模块；针对服务提供商来说，视频检索技术根据应用模式不同可分为通用视频检索、特定视频检索以及视频主动推荐；而针对终端用户来说，视频检索技术根据查询输入的不同可分为基于文本关键词、中层语义概念和视频样例的检索，以及不需要查询的视频主动推荐技术。

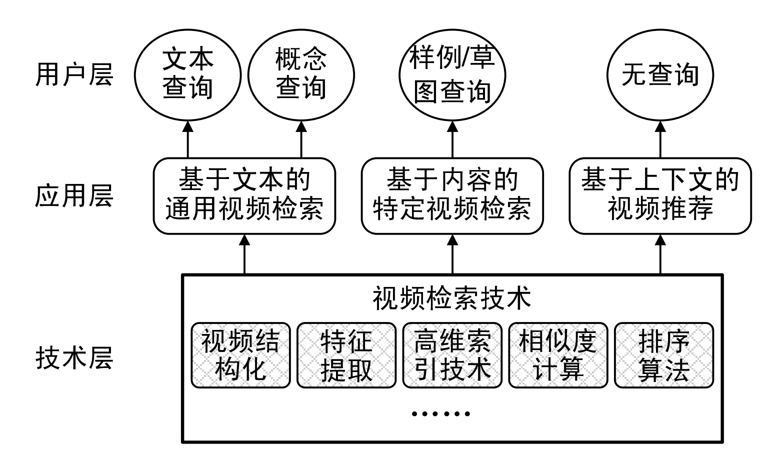


图 1. 视频检索技术在不同层次的表现形式

现有的商业化视频搜索引擎，如百度，Google Video，Blinkx 等主要依赖文本检索技术，通过从视频元数据中提取视频标题、描述、标签、字幕文本、语音识别文本等信息进行基于文本的视频检索，用户查询入口为文本关键词。这类方法在视频文本缺失（如家庭视频等），以及视频文本不能准确描述视频内容（如文本标签错误）时，检索性能大大降低。

因此，从 90 年代开始，基于内容的视频检索(Content-Based Video Retrieval, CBVR)技术应运而生[1][2][3]。这类方法直接从视频本身提取底层视觉特征进行索引和相似度计算，支持用户基于示例的检索(Example-based retrieval)和基于草图的检索(Sketch-based retrieval)。目

前，基于内容的视频检索方法还无法应用于通用视频检索，仅在一些小型的实验系统中使用，如：IBM 开发的 QBIC 检索系统[4]、意大利帕勒莫大学(Università degli Studi di Palermo)开发的 JACOB 系统[5]、美国哥伦比亚大学开发的 VideoQ 视频查询系统[6]和网络视频搜索引擎 WEbSEEk[7]等。**值得注意的是，在某些特定领域，基于内容的视频检索已体现出重要的应用价值**，如版权保护中非法拷贝视频检测、大规模网络视频中的重复视频检测、监控视频中的特定语义事件检测等。以视频拷贝检测技术为例，由于其具有重要的应用需求和价值，由美国国家信息标准协会举办的国际视频检索评测（TRECVID）[13] 从 2008 年开始，设立了一项“视频拷贝检测评测”任务。通过逐年的评测，目前该领域已取得突破性的进展[55][9]。

基于内容的视频检索面临的本质问题是“语义鸿沟”（Semantic Gap）。斯穆德斯（Smeulders）等[8]将该问题定义为“机器从视频中提取的底层特征和用户所理解的高层语义之间缺少一一对应关系”。为了缩小语义鸿沟，近年来，多媒体领域出现了一个非常有前景的研究方向——基于概念的视频检索[10][11][12]。这类方法在视频底层特征描述与用户语义查询之间引入一个中间语义概念层，包含的概念具有一定语义，同时又可以基于底层特征训练概念检测器，通过机器自动识别，如：物体对象类概念（人、飞机、山、路、船、建筑物等），场景类概念（室内/室外、水景、雪景、沙漠等），事件类概念（起飞、运动、行走）等。通过分别建立从低层特征到语义概念（即语义概念检测）和从用户查询到语义概念（即查询分析）的两层映射，最终实现基于概念的语义视频检索。从 TRECVID 近三年的评测结果可以发现，该方法的性能远远高于单纯基于文本或者视觉的视频检索方法。

近些年，随着 Web2.0 技术的发展，大部分视频数据主要通过网络平台进行存储和传播，如 YouTube、土豆网、优酷网等。这些平台为视频数据提供了一个可供用户交流的网络环境（社会网络，social network[1]）。除了视频本身的相关性之外，丰富的上下文信息也为视频之间建立了连接，如：同一个作者上传的视频之间具有一定相似性，而被同一个用户评论过的视频之间也具有一定关联。杰恩(R. Jain)[16] 和 靳辛(音译，X. Jin)[17]等在 ACM Multimedia 2010“brave new idea”中强烈呼吁大家使用上下文信息进行多媒体内容分析。首先，不考虑上下文的内容是没有意义的[16]，如不同用户对同一个视频会有不同的理解和标注，即使是同一个用户，在不同的时间段对同一个视频的理解也是不同的。其次，丰富的上下文信息对于克服网络多媒体数据特征稀疏、噪声大等问题具有重要的现实意义。因此，近两年来，上下文信息逐渐受到多媒体研究人员的关注，并在图像推荐和预测[18][21]、视频分类[20]、视频话题发现[19][57]等领域出现了一些尝试性的工作。

综上所述，视频检索技术的发展经历了一个从文本、视觉内容、语义概念，到上下文信息的发展过程。由于基于文本的视频检索主要采用已有的文本信息检索技术，因此本文不再介绍。在后续章节，本文将分别以基于内容的视频拷贝检测技术、基于概念的视频检索技术以及基于上下文的视频话题挖掘和推荐技术为例，对视频检索研究现状进行介绍。最后，本文也对本课题组在相关研究上所取得的进展进行了简要介绍。

2 基于内容的视频拷贝检测技术

美国国家标准局将**视频拷贝**定义为：一个视频或者其片段在经过某些编辑处理后，得到的内容相同但视觉外观(如亮度)不完全一致的同源视频版本[13]。视频拷贝检测（Copy Detection）即指通过将查询视频的内容特征与库中视频做匹配，判断此查询视频是否是库中某个源视频的拷贝。不同于视频检索，被拷贝的视频在源视频的基础上进行了各种几何和图

[1] 亦有译作“社交网络”或“社区网络”

像变换，使得视频在视觉上发生了不同程度改变，称为**拷贝攻击**[51]。常见的拷贝攻击有编码方式转换、画面尺寸变化、画面比例变化、添加边框等。

根据使用特征的不同，已有基于内容的拷贝检测方法可以分为三大类[55]：基于数字签名的方法，基于关键帧的方法，以及基于轨迹的方法。

基于数字签名的方法通常将整个视频内容表示为一个全局的特征值，从而进行视频级的快速匹配。如将视频里所有帧的颜色直方图[24]，排序特征[52]等，进行平均。文献[51]已经证明，这类方法只对较小的拷贝攻击有效。而且由于忽略了视频的时序信息，因此只能检测针对整个视频的拷贝，无法识别部分片段的拷贝。

基于关键帧的方法从视频中抽样出具有代表性的的帧进行匹配。其中核心算法是如何对两个不等长的关键帧序列进行匹配。如邱志义（C.Y. Chiu ）等在文献[54]中使用动态规划算法选取最长的匹配序列；吴晓（X. Wu）等在文献[24]中采用滑动窗口的方法进行关键帧序列匹配；陈汉勤（音译，Hung-Khoon Tan）等在文献[55]中将帧之间的时序关系表示成一个有向边的时序网络（Temporal Network），在帧级匹配的结果上，基于时序网络进行视觉-时序一致性验证，准确检测和定位视频拷贝片段。这类方法利用的时序关系对视觉变化有一定的鲁棒性，但对时域的变化，如前后片段调换，帧率改变和丢失帧等非常敏感。

基于轨迹的方法通过跟踪兴趣点在视频序列中的变化，形成具有时-空（spatio-temporal）信息的轨迹特征。如 J. Law-To 等在文献[53]中利用轨迹特征标注不同的运动行为，吴晓等人在文献[9]中采用轨迹词袋的方法解决不连续的时序模式问题。轨迹特征同时考虑了兴趣点在空间和时序上变化，对复杂的拷贝攻击具有鲁棒性，但由于提取兴趣点和轨迹非常耗时，所以这类方法的时间复杂度比较高。

3　基于概念的语义视频检索技术

基于概念的视频检索框架如图 2 所示，包含三个关键步骤：一是语义概念集的建立，即选取多少个概念，以及选取哪些概念构建语义空间；二是建立从低层特征到语义概念的映射关系，即语义概念检测；三是建立用户查询到语义概念的映射关系，即查询分析。下面我们将分别介绍这三个方面的研究现状。

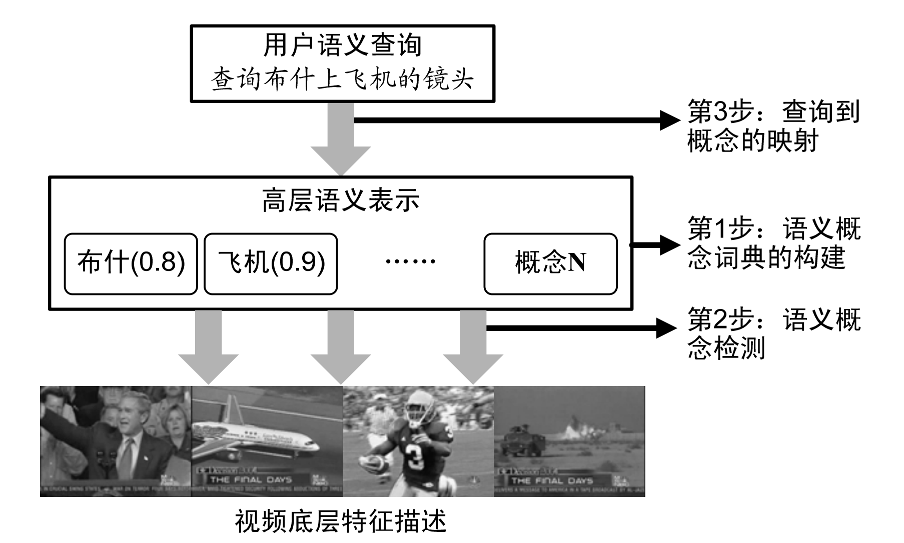


图 2.　基于概念的视频检索框架

3.1 语义概念集构建

基于概念的视频检索第一步就是定义一个合适的语义概念集。目前被广泛认可的概念集包括 LSCOM（Large-Scale Concept Ontology for Multimedia）[14]和 Mediamill-101[15]。LSCOM 是由美国 IBM 沃森研究中心、卡内基.梅隆大学（CMU）和哥伦比亚大学联合开发，包含 2000 个语义概念的定义，并在 TRECVID 2005 视频集上对 449 个概念进行了人工标注，为多媒体检索方法的研究提供了重要数据集。在 LSCOM 的基础上，研究者进一步精选了 44 个概念，构成 LSCOM-lite 词典。将语义空间划分成七个相互正交的子空间：对象（objects）、行为（activities）、事件（events）、场景(scenes/locations)、人物（people）、图表（graphics）和节目（program），并根据查询中概念单词的使用情况为每个子空间选择合适的概念。Mediamill-101 是由阿姆斯特丹大学开发，并在 TRECVID2005 的视频数据集上进行了人工标注。

基于这些语义概念词典，美国卡内基.梅隆大学 豪普特曼（A. Hauptmann）等对语义概念集的构建进行了一系列基础性的研究[11]，得到一个重要结论：当概念集的规模在 5000 左右，每个概念的检测精度不低于 10%的情况下，基于概念的视频检索可以达到与文本检索相当的效果（MAP[2]=65%）。该结论为后续基于语义概念的视频检索技术发展奠定了基础。2008 年，美国德克萨斯大学的卢亦娟（音译，Y. j. Lu）等[28]进一步提出：不同的概念，具有不同大小的语义鸿沟。如概念“Sunset”很容易用视觉特征来描述，语义鸿沟较小，而“Europe”很难用简单的视觉特征描述，具有很大的语义鸿沟。语义鸿沟较小的概念相对容易通过底层特征实现机器自动检测，适合构造语义概念词典。基于以上理论，他们首次提出对语义鸿沟进行量化，从而自动选择语义鸿沟最小的概念构建语义概念集。这种方法无需人工干预，具有很强的操作性。

这些大规模的、标准的、带标注数据的语义概念集的建立以及构建理论的不断完善，对于提高视频检索精度，规范视频检索评测具有重要意义。

3.2 语义概念检测

对于上述定义的每一个概念，需要通过机器学习的方法从已标注的正负样本中学习来建立概念检测器。在过去的十年，针对视频/图像的语义概念检测得到了广泛的研究[40]。其核心模块包括：视频底层特征提取、学习模型以及多模态的特征融合。

首先，有效的特征表示是概念检测成功的关键。颜色（如颜色直方图、颜色矩）和纹理特征（如小波纹理等）是计算机视觉里被普遍使用的两类视觉特征。与这些描述视频/图像整体分布特性的全局特征相比，局部特征对图像的几何和光照变化等具有鲁棒性，近年来，在很多视觉分类任务中显示了突出的效果。局部特征的提取包括局部特征点检测与描述两部分。目前被广泛采用的特征点检测方法包括哈里斯（Harris）角点检测算法[30]和高斯差分（difference of gaussian， DoG）的局部特征检测方法[29]等，描述方法如罗伊（Lowe）提出的尺度不变特征转换（Scale-invariant feature transform， SIFT）局部特征描述子[29]。关于局部特征点检测和描述的详细综述可见文献[31]和[32]。由于每幅图像提取的局部特征点数量庞大，所以不能直接用来描述视觉内容。典型的局部特征使用方法是视觉词典（visual vocabulary）。首先将局部特征点聚类成视觉词（visual word），从而产生视觉词典。其次，将每个图像的局部特征点映射到视觉词典，得到每个图像的视觉词向量表示（bag-of-visual-words， BoW）[33]。

2 Mean Average Precision，系统平均准确率

基于上述特征，可以为每个概念通过学习建立分类器。目前在概念检测任务中广泛使用的学习模型是支持向量机（Support Vector Machine）[34]。其次还有高斯混合模型（Gaussian mixture models）[35]，隐马尔可夫模型（Hidden Markov models）[36]等。上述方法都是针对单个语义概念建立模型。但视频中语义概念并不独立存在，不同的语义概念之间往往存在一些上下文（Context）约束或者共现(Co-occurrence)关系。比如检测到“天空”和“绿地”会增加检测到“风景”的概率，而减少检测到“室内”的概率。因此，还需要研究利用不同概念之间的相互关系，增强或排除一些概念。针对这一现象有学者提出了基于多概念的视频语义表示方法。比较有代表性的方法包括：豪普特曼等[38]提出的通过对数回归获取概念间关系、进行多概念融合的方法；以及清华大学在 TRECVID2007 中提出的基于贝叶斯-狄里克利度量（Bayesian Dirichelet Metric）和神经网络的方法[39]等。

多模态融合方法包括前融合和后融合。前融合是指将各种特征组合成一个长的特征向量，基于该向量训练一个概念检测器；而后融合是指为每个特征训练一个概念检测器，通过融合每个检测器的输出结果作为最终的检测结果。两者各有利弊。前者隐含地考虑了不同特征之间的互补关系，然而需要面对高维特征处理的问题；后者相对容易实现，并被很多概念检测系统采用，但如何对多个检测器进行加权是关键。斯诺克（Snoek）等[37]对这两种融合方法进行了分析比较，并提出了一个自适应的多模态特征选择算法。

3.3 概念映射

基于上述方法为每个语义概念构建分类器后，通过将用户查询映射到相关的概念检测器，则可以实现基于概念的视频检索。根据使用的特征不同，这种检索可以分为：基于文本特征的映射、基于视觉内容的映射和基于反馈结果的映射。

由于文本是对视频语义内容最直接的描述，所以目前大部分系统都采用文本特征将查询映射到语义概念[27][41][42][43]。一种方法是基于知识本体的概念映射，如 WordNet[45]等。这些知识库通常包含词之间的关系结构，如上位关系、下位关系、同义词关系等，以及词之间的语义相似度度量算法，如：RES[46]、Lesk[50]、WUP[47]、JCN[48]以及近期提出的 OSS[44]方法等。基于这些度量方法，可以实现查询关键词到语义概念之间的映射。另一种是数据驱动的概念映射。这类方法通过统计模型，如潜在语义分析（Latent Semantic Indexing）[49]等，分析数据库中各个词项之间的共现情况，从而自动挖掘词项之间的相关性。

除查询文本外，有时查询会以图像样例或者视频片段等视觉形式给出。因此基于这些视觉内容也可以完成查询到概念之间的映射。其一般流程是：将上一节介绍的概念检测器对查询样例进行对应的概念检测，然后直接选择后验概率较高的概念作为该查询的概念映射结果[58]。由于这类方法对概念检测器的检测精度非常敏感，一旦检测器对查询样例判别错误，则直接导致概念映射错误，因此，研究者们提出把这些带有噪声的概念检测结果作为特征，在此基础上进行进一步的统计分析[61]或者机器学习[42]，得到更稳定的概念映射结果。

与基于整个数据集的统计分析相比，研究人员认为，在与查询相关的一个特定子集中统计查询与概念之间的相关性更具有价值。通常这个与查询相关的特定子集需要用户标注产生，称为相关反馈（Relevance Feedback）。该方法在用户标注的集合中提取特征，用上述基于文本特征的方法，或基于视觉特征的方法，将查询映射到语义概念。为了减少用户的参与，有些系统简化标注环节，默认初始检索结果中前 N 个结果为正样本，最后 M 个结果为负样本，称为伪相关反馈（Pseudo-Relevance Feedback）。由于伪相关反馈方法依赖初始的检索结果，所以，在初始结果很差的情况，伪相关反馈方法会降低检索性能[12]。

4 基于上下文的网络视频分析技术

上下文是指某个对象存在或发生所依赖的条件和环境[16]。只有考虑上下文信息，才能正确理解视频包含的语义内容，有效缩小语义鸿沟；同时，上下文信息可以有效地缩小检索空间，提高检索性能。如一个在澳大利亚拍摄的视频，不太可能出现雪景。具体到网络视频，我们可将其分为以视频为中心的上下文，包括视频属性，如长度、类别等；拍摄设备的参数，如摄像机型号、分辨率等；拍摄环境，如拍摄地点、时间等；以及以用户为中心的上下文，如用户对视频的标注、评论、收藏等网络行为所产生的社会网络。下面我们将分别介绍两类上下文信息在网络视频内容分析中的研究现状。

杰恩等[16]采用照相机的 EXIF[3]参数进行图像分类，得到了比基于内容更好的效果。吴晓等人[24]通过考虑视频的时间长度信息提高近似视频检测的准确率。随着 GPS 设备的普及，视频地理信息的价值不断被发现。文献[19]和[26]分别提出了一个 GeoFolk 的框架，以及一个潜在地理性话题分析（Latent Geographical Topic Analysis， LGTA）方法，基于视频的地理信息发现具有地域性的视频话题，用于比较同一个话题在不同地区的发展，以及不同地区的热点话题对比等。

另一方面，网络用户行为所产生的上下文信息包含丰富的统计知识[17]。本维奴托（Benevenuto）等[25]对 YouTube 用户的视频回复行为进行深入分析，得到了多个有价值的统计模型，可用于后续网络视频分析；罗劳夫（Roelof）等[18]根据用户的订阅行为预测每个用户最喜爱的照片，在 Flicker 获取的数据集上的实验显示，基于上下文信息的平均预测精度（92%）分别高于文本（87%）和视觉（88%）；吴晓等基于 YouTube 网站提供的相关视频的类别信息进行投票，实现视频自动分类；苟良（L. Gou）等[23]提出了一种社会网络文本排序算法（Social Network Document Rank, SNDorRank），通过计算查询用户的网络与视频作者的网络之间的相关性，对视频检索结果进行排序，实现更贴近用户兴趣的视频检索。

5 我们的工作

近三年来，本课题组在视频内容分析研究方面取得了很多进展，并开发了多个系统。本节将重点介绍我们在大规模网络视频拷贝检测、基于概念的网络视频检索以及基于上下文的网络视频话题发现与检索三个方面的研究进展和开发的相关系统。

5.1 大规模网络视频拷贝检测系统

在视频拷贝检测方面，我们以提高检测精度与检测效率为目标，提出了多种基于单帧的和基于视频的拷贝检测特征，并尝试通过高维索引技术，GPU 加速等技术来提高检索效率。开发的视频拷贝检测系统分别在 2008 年和 2009 年视频检索国际评测（TRECVID）的视频拷贝检测项目中，分别获得总成绩第三名和第一名的好成绩[66]。

5.1.1 面向复杂攻击的鲁棒视觉特征挖掘方法

各种复杂的视频拷贝攻击对视觉特征提出了苛刻的要求。我们提出了融合样例自动扩展与稳定特征挖掘的高鲁棒性视觉特征提取理论与方法[68]该方法引入全仿射空间概念，通过自动模拟不同视角下的图像仿射形变情况，将原有特征扩展到图像在不同仿射条件下检测到

[3] Exchangeable image file format，可交换图像文件格式。实际上 EXIF 格式就是在 JPEG 格式头部插入了数码照片的信息，包括：拍摄时的光圈、快门、白平衡、ISO、焦距、日期时间等各种拍摄条件以及相机品牌、型号、色彩编码、拍摄时录制的声音和全球定位系统（GPS）、缩略图等。

的局部特征集合。其次，为了从这些大量的扩展信息中找到最具稳定性的代表性特征，我们采用基于全局稳定度的稳定特征挖掘方法来得到各图像中所有具有高鲁棒性的局部特征集合，以仅占扩展信息 5%的特征作为图像在各种复杂攻击方式下的视觉信息表征。

与普通的图片/视频检索不同，拷贝图片/视频都是经过拷贝攻击处理的，如何度量这种经过各种变换后的图片之间的相似性是拷贝检测的核心问题。在最近的工作中，我们提出了一种基于匹配对的几何一致性度量方法[56]。与传统的直接计算两个匹配到的关键点之间的相似度不同，该方法通过计算两两匹配对的几何变换之间的相似性，来度量两幅图像之间的几何一致性。具有相似变换的匹配对越多，说明两幅图像越有可能是拷贝。这个方法的优势是既能处理全局的拷贝攻击，如缩放，旋转，位移等，也能处理局部变换，以及一定程度的视角扭曲。其次，能同时处理一幅图像中存在的多种视觉模式变换。

5.1.2 面向高速匹配的高维特征索引技术

由于局部特征的个数和维数都远远超出了传统匹配方法的应对能力，因此为特征建立有效的高维索引是实现大规模网络视频拷贝检测的必要环节。我们提出了一种面向非均匀数据分布的局部敏感哈希(Locality Sensitive Hashing，LSH)高维索引方法[67]，该方法利用数据分布信息来选择投影向量，即通过非监督学习的方法获得投影向量。同时，为了直观地分析哈希函数的性能，我们提出了数据分布熵的概念。通过评估数据分布熵来选择较优的哈希函数。这样产生的哈希函数，在尽量保留原始数据近邻关系的情况下，使得各哈希表项索引的数据更均匀。通过在著名的开放数据库上进行验证可以看出，在相同精度下，我们的索引算法比原始的 LSH 算法减小了 30%的内存消耗。同时，在使用相同个数的哈希表时，查询精度和效率都有提高。

5.2 基于概念的语义视频检索系统

本课题组从 2007 年开始，一直从事基于概念的通用视频检索研究，并取得了重要成果。我们研发的基于隐含语义概念的视频检索系统在国际视频检索评测(TRECVID)中，分别获得 2007 年自动检索任务第二名[27]，2008 年第一名[60]，以及 2009 年交互式检索任务第二名[65]。下面将对两个概念选择算法，以及两个隐含语义与显性语义融合算法进行介绍。

5.2.1 多模态的概念选择方法

除了考虑查询与概念之间的语义相似性外，不同概念在检索中扮演着不同的角色。 例如，对于查询 “Find shots of one or more people at a table or desk, with a computer visible.”来说，虽然概念“Face”和“Person”与查询很相关，但由于这两个概念在正负样本中的分布很类似，所以对于正负样本没有区分能力；另一方面，概念“Computer”和“Hand”和查询很相关，且具有很强的区分能力，但由于这两个概念的机器自动检测精度很低，因此对于检索贡献不大。基于上述分析，我们提出了一种基于分布的概念选择方法（Distribution Based ·Concept Selection, DBCS）[61]。通过融合概念检测器的可信度，以及概念分布在相关集和不相关集中的可区分性来选择最有价值的概念进行查询。

为了考虑查询文本描述不完整的问题，在此基础上，我们进一步提出了一种基于多模态概念关联图的概念选择模型[62]，将查询与概念之间的关系表示成一个网状的关联图，分别包含查询与概念之间，以及概念与概念之间的相似关系，同时支持查询样例与查询文本到语义概念之间的多模态映射。通过流行排序算法，将查询与概念之间的多模态相似性在整个关联图中进行传播，直到网络达到稳态，从而选择相似度最大的前 N 个概念进行视频检索。与多种基于星型结构的概念视频检索方法比较，该方法针对查询文本比较稀疏的情况具有较

强的鲁棒性，平均精度提高了近 20%。

5.2.2 显性语义概念与隐含语义概念融合的视频检索系统

目前基于概念的视频检索需要人工定义一个有限的概念集（本文称为显性语义）。由于该概念集无法覆盖整个查询语义空间，在实际检索过程中会出现零概率映射和不可扩展等问题。其次，学习概念检测器需要人工标注大量的训练数据，费时费力。因此，研究人员开始尝试新的解决办法，试图通过概率主题模型，无监督地从视频底层特征中提取隐含主题（本文称为隐含语义）。我们提出了一个隐含语义和显性语义相结合的语义视频检索框架[60]，通过隐含狄利克雷分配（Latent Dirichlet Allocation，LDA）模型从底层特征描述中提取具有稳定性的特定的隐含语义；同时基于上述概念选择算法，将用户查询映射到人工定义的显性语义概念集，融合两种概念来实现视频检索。通过隐含语义的数据驱动特性来弥补显性语义检索中的零概率映射问题，提高检索召回率，同时通过查询到固定显性语义概念集的准确映射，保证检索精度。在此基础上，我们进一步提出了基于二分图的融合算法[63]，根据查询的不同，对两种概念进行自适应的加权融合。

5.3 基于上下文的大规模网络视频分析

我们在基于上下文的大规模网络视频话题自动发现和推荐方面取得了重要进展，同时在基于多种上下文信息的视频检索方面取得了探索性的成果。下面将分别介绍这两个内容。

5.3.1 基于轨迹的网络视频话题发现与推荐

根据 YouTube Report 2009[22]的统计，有 45%的用户登录 YouTube 并没有明确的检索目标，而是浏览网站主动推荐的“热点视频”和“热点话题”，表明这种不需要用户输入查询的视频话题自动发现和推荐模式越来越受到网络用户的欢迎。为了提高网络视频特征的可靠性，我们提出了一种基于全局轨迹特征的网络视频话题检测方法[59]。首先，将每个标签（tag）表示为时间轴上的特征轨迹，仅从轨迹中提取显著点（轨迹中的顶点）进行聚类，产生发生在该时间点的事件。这种考虑上下文的轨迹特征能有效过滤噪声。其次，通过计算事件之间的文本相似度和视觉拷贝检测相似度，建立事件发展链接图。通过在图上寻找最优路径，提取最热门的前 N 个话题轨迹。该方法通过考虑全局的链接情况来判断这些事件是否构成一个话题，因此对于局部的错误链接具有较强的鲁棒性。此外，上述这种基于轨迹的话题发现方法，除了能发现内容热点话题之外（content-hot），还可以发现轨迹发展热点话题（evolution-hot）。通常这类话题都是在互联网上具有争议性的内容，在一段时间内被反复讨论；另一类是潜在热点话题（potential-hot），这类话题目前仅被少数人群关注，但有不断发展的趋势，很有可能在后续某个点爆发。后续两类话题是传统的基于内容的方法无法发现的，但他们在网络监管中具有重要意义。综合上述方法，我们实现了一个基于轨迹的网络视频话题自动发现和展示系统[57]，具有很好的用户体验。

5.3.2 基于社会信息的网络视频检索

不同的上下文信息之间具有一定的关联，如同一个作者上传的不同视频更有可能被相同的用户评论。而目前已有的方法大都局限于对一种信息的研究。我们提出了一种基于社区结构重排序的视频检索方法[64]，将用户之间、视频之间以及用户和视频之间的多种链接关系形式化为一个异构的上下文网络，通过从该网络中提取隐含的社区结构（community），挖掘多种上下文信息之间稳定的关联模式，实现基于社区结构的视频检索结果重排序。在包含 82352 个 YouTube 视频和 39555 个用户的异构网络中进行实验比较，该方法的检索结果均优于基于纯文本和纯视觉的方法。

6 总结

基于内容的视频检索技术经历了近十年的发展，尽管在某些特定领域，如视频拷贝检测，取得了重要进展，但目前的技术水平还不能满足用户对通用视频进行基于内容的检索需求。其中的技术瓶颈主要是语义鸿沟问题。因而，目前商用的通用视频检索主要还是基于文本信息检索技术。但是，随着视频网站的普及，带有丰富上下文信息的网络视频数据已成为视频检索的主要对象。这些上下文信息为我们绕开复杂的视频内容本身，从视频所处的上下文环境出发去缩小语义鸿沟提供了一种可能，使得基于上下文信息的网络视频分析和检索成为当今网络多媒体时代的研究热点。同时，我们也相信视频检索技术和网络视频数据的结合会催生更加丰富的网络多媒体应用。

参考文献：

[1] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State-of-the-art and challenges. In *ACM Transactions on Multimedia Computing, Communication, and Applications*, 2006.

[2] C.W. Ngo, H.J. Zhang, and T.C. Pone, Recent Advances in Content Based Video Analysis, *International Journal of Image and Graphics*, December 2001.

[3] N. Dimitrova, H.J. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, Applications of Video-Content Analysis and Retrieval, *IEEE Multimedia*, vol. 9, no.3, pp. 42-55, 2002.

[4] M. Flickner, H. S. Sawhney, et al .Query by image and video content: the QBIC system. *IEEE computer*, 28(9):23-32, 1995.

[5] L. Marco, A. Edoardo, JACOB: Just a content-based query system for video databases. *Proc. ICASSP* , Atlanta, G A, 1996

[6] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong, VideoQ: An Automatic Content-Based Video Search System Using Visual Cues, *ACM Multimedia*, Seattle, WA, November 1997.

[7] J. R. Smith and S. F. Chang, Visually searching the web for content. *IEEE Multimedia* , pp.12–20, 1997.

[8] A. W. M. Smeulders, M.Worring, S. Santini, A. Gupta, and R. Jain, Content based image retrieval at the end of the early years, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1349–1380, 2000.

[9] X. Wu, M Takimoto, J Adachi. Scene duplicate detection based on the pattern of discontinuities in feature point trajectories. 、*ACM international conference on Multimedia*, pp.51-60, 2008

[10] C. G. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322, 2009.

[11] A. Hauptmann, Y. Rong, W.H. Lin, M. Christel, H.Wactlar, Can High-Level Concepts Fill the Semantic Gap in Video Retrieval? A Case Study With Broadcast News, *IEEE Transactions on Multimedia*, 9(5): 958-966, 2007

[12] A. Natsev, A. Haubold, J. Teˇsi´c, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval: A comparative review and new approaches. In *ACM Multimedia (ACM MM)*, 2007.

[13] A. F. Smeaton, P. Over, and W. Kraaij, Evaluation campaigns and TRECVID, *in Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 321–330, 2006.

[14] L. S. Kennedy, Revision of LSCOM event/activity annotations, *Technical Report 221-2006-7,*

Columbia University ADVENT Technical Report, 2006.

[15] C. G. M. Snoek, M. Worring, J. C. v. Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM Multimedia*, 2006.

[16] R. Jain and P. Sinha. Content without context is meaningless. In *ACM Multimedia (ACM MM)*, pp.1259-1268, 2010.

[17] X. Jin, A. Gallagher, L. L. Cao, J. B. Luo, and J. W. Han. The wisdom of social multimedia: using flickr for prediction and forecast. In *ACM Multimedia* (ACM MM '10), pp.1235-1244, 2010.

[18] V. Z. Roelof, R. Adam, and G. P. Lluis. Prediction of favourite photos using social, visual, and textual signals. *In ACM Multimedia*, pp.1015-1018, 2010.

[19] S. Sizov. Geofolk: Latent spatial semantics in web 2.0 social media. In *ACM International Conference on Web Search and Data Mining*, 2010.

[20] X. Wu, W. L. Zhao, and C. -W. Ngo. 2009. Towards Google challenge: combining contextual and social information for web video categorization. In *ACM Multimedia* (MM '09), pp.1109-1110, 2009.

[21] R. R. Ji, X. Xie, H. X. Yao, W. Y. Ma: Mining city landmarks from blogs by graph modeling. *ACM Multimedia*, pp.105-114, 2009.

[22] YouTube report 2009, http://youtubereport2009.com/.

[23] L. Gou, H. H. Chen, J. H. Kim, X. Zhang, SNDocRank: a Social Network-Based Video Search Ranking Framework. *In ACM MIR* ,2010.

[24] X. Wu, C.-W. Ngo, A. G. Hauptmann and H. K. Tan. Real-Time Near-Duplicate Elimination for Web Video Search with Content and Context. *IEEE Transactions on Multimedia*, 11(2): 196-207, 2009

[25] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross. Video interactions in online video social networks. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5:30:1-30:25, 2009.

[26] Z. J. Yin, L. L. Cao, J. W. Han, C. X. Zhai, and T. Huang. Geographical topic discovery and comparison. *In international conference on World Wide Web*, pp. 247-256, 2011.

[27] S. Tang, Y. Zhang, J. Li, J. Cao, H. Luan, Q. He, and X. Zhang，TRECVID 2007 Search Tasks by NUS-ICT，*In NIST TRECVID Video Retrieval Workshop*, 2007.

[28] Y. J. Lu, L. Zhang, Q. Tian, W. Y. Ma, What Are the High-Level Concepts with Small Semantic Gaps, *CVPR* ,2008.

[29] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints[J]. *International Journal of Computer Vision*, 2004.

[30] C. Harris, M. Stephens. A combined corner and edge detector, *In. Proceeding of 4th Alvey Vision Conference*, pp.147-151, 1988.

[31] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool. A comparison of affine region detectors, *In IJCV* , 65(1/2):43-72, 2005.

[32] K.Mikolajczyk, and C.Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[33] J. Sivic, A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos, *International Conference on Computer Vision*, 2003.

[34] J. Tesic, A. Natssev, and J.R. Smith. Cluster-based data modeling for semantic video search. *ACM International Conference on Image and Video Retrieval (ACM CIVR)*, 2007.

[35] A. Amir, W.H., G. Iyengar, C.-Y.Lin, M. Naphade, A. Natsev, C. Neti, H. J. Nock, J. R. Smith, B. L. Tseng, Y. Wu, and D. Zhang. IBM research TRECVID-2003 video retrieval system. *in NIST*

TRECVID, 2003.

[36] B. Pytlik, A.G., D. Karakos, and S. Khudanpur. Trecvid 2005 experiment at johns hopkins university: Using hidden markov models for video retrieval. *in NIST TRECVID*, 2005.

[37] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. *In ACM Multimedia (ACM MM)*. 399-402, 2005.

[38] A. Hauptmann, M.-Y.C. et al., Confounded expectations: Informedia at TRECVID 2004, *NIST TRECVID Workshop*, 2004.

[39] J. Yuan, Z. Guo, THU and ICRC at TRECVID 2007, *NIST TRECVID 2007 Workshop*, USA , Nov. 2007.

[40] Y. G. Jiang, J. Yang, C. W. Ngo, and A. G. Hauptmann. Representations of Keypoint-Based semantic concept detection: A comprehensive study, IEEE Transactions on Multimedia, 12(1): 42-53, 2009.

[41] A. Natsev, A. Haubold, J. Teˇsi´c, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval: A comparative review and new approaches. *In ACM Multimedia (ACM MM)*, 2007.

[42] M. Campbell, A. Haubold, M. Liu, A. P. Natsev, J. R. Smith, J. Tešić, L. Xie, R. Yan and J. Yang，IBM Research TRECVID-2007 Video Retrieval System， *In NIST TRECVID Video Retrieval Workshop*, 2007.

[43] T. Mei , X. Hua and et al. MSRA-USTC-SJTU at TRECVID 2007: High-level feature extraction and search, *In NIST TRECVID Video Retrieval Workshop*. 2007.

[44] X.-Y. Wei, C.-W. Ngo, Y.-G. Jiang, Selection of Concept Detectors for Video Search by Ontology-Enriched Semantic Spaces, *IEEE Transaction on Multimedia*, Vol. 10, no. 6, 2008.

[45] C. Fellbaum and Ed. WordNet: an electronic lexical database. The MIT Press, 1998.

[46] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy. *In IJCAL*, pp.448-453, 1995.

[47] Z. Wu and M. Palmer. Verb semantic and lexical selection. *In Annual Meeting of the ACL*, pp.133–138, 1994.

[48] J. J. Jiang. Semantic similarity based on corpus statistics and lexical taxonomy. *In ROCLING*, 1997.

[49] T. Hofmann. Probabilistic latent semantic indexing. *In Proc. of the 22nd Intl. ACM SIGIR conference*, pp. 50–57, 1999.

[50] M. E. Lesk, Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, *SIGDOC Conference*, Toronto, Ontario, June, 1986.

[51] J. Law-To, L. Chen, A. Joly, et al. Video copy detection: a comparative study. Proceedings of the 6th ACM international conference on Image and video retrieval, pp.371 – 378, 2007.

[52] X.S. Hua, X. Chen, H. J. Zhang. Robust video signature based on ordinal measure, *International Conference on Image Processing*, 2004.

[53] J. Law-To, O. Buisson, V. Gouet-Brunetand, and N. Boujemaa. Robust voting algorithm based on labels of behavior for video copy detection, *ACM International Conference on Multimedia*, pp.835-844, 2006.

[54] C. Y. Chiu, C. H. Li, H. A. Wang, et al. A time warping based approach for video copy detection[C], *Proceedings of International Conference on Pattern Recognition*, Hong Kong, pp.228-231, 2006.

[55] H. K. Tan, C. W. Ngo, and T. S. Chua, Efficient mining of multiple partial near-duplicate alignments by temporal network, *IEEE Transactions on Circuits and Systems for Video Technology(CSVT)*, vol. 20, no. 11, pp. 1486–1498, 2010.

[56] H. T. Xie, K. Gao, Y. D. Zhang, S. Tang, J. T. Li, Efficient Feature Detection and Effective Post-Verification for Large Scale Near-Duplicate Image Search， *IEEE Transaction on Multimedia*, accepted, 2011.

[57] J. Cao, C.-W Ngo, Y. D. Zhang, L. Ma, Trajectory-based Visualization of Web Video Topics, *ACM International Conference on Multimedia (ACM MM)*, Florence, Italy, 2010.

[58] X. Li, D. Wang, J. Li, B. Zhang: Video Search in Concept Subspace: A Text-Like Paradigm, *ACM International Conference on Image and Video Retrieval*, pp.603-610, Amsterdam, the Netherlands, 2007

[59] J. Cao, C.-W. Ngo, Y. D. Zhang, J. T. Li, Trajectory-based Visualization of Web Video Topics, *IEEE Transactions on Circuits and Systems for Video Technology(CSVT)*, 2011.

[60] J. Cao, Y. D. Zhang, B. L. Feng, X. F. Hua, L. Bao, X. Zhang and J. T. Li , TRECVID 2008 Search Task by MCG-ICT-CAS，In *NIST TRECVID Video Retrieval Workshop*. 2008.

[61] J. Cao, H. F. Jing, C.-W. Ngo, Y. D. Zhang, Distribution-based Concept Selection for Concept-based Video Retrieval, *ACM International Conference on Multimedia (ACM MM)*, Beijing, China, Oct. 2009.

[62] B. L. Feng, J. Cao, L. Bao, Y. D. Zhang, S. X. Lin, X. G. Bao, X. C. Yun. Graph-Based Multi-Space Semantic Correlation Diffusion for Video Retrieval. *International Journal of Visual Computer*, in press, 2011.

[63] L. Bao, J. Cao, Y. D. Zhang, M. Y. Chen, J. T. Li, A. Hauptmann, Explicit and Implicit Concept-based Video Retrieval with Bipartite Graph Propagation Model, *ACM Multimedia*, Florence, Italy, 2010.

[64] L. Pang, J. Cao, Y. D. Zhang, S. X. Lin, Leveraging Collective Wisdom for Web Video Retrieval through Heterogeneous Community Discovery , *ACM MM 11*, accepted , 2011.

[65] J. Cao, Y. D. Zhang, B. L. Feng, L. Bao, L. Pang and J. T. Li ,TRECVID 2009 Interactive Search Task by MCG-ICT-CAS，*In NIST TRECVID Video Retrieval Workshop*. 2009.

[66] K. Gao, X. Wu, H.-T. Xie, W. Zhang, Z.-D. Mao, TRECVID 2009 Copy Detection Task by MCG-ICT-CAS, *In NIST TRECVID Video Retrieval Workshop*. 2009.

[67] W. Zhang, K. Gao, Y. D. Zhang, J. T. Li, Data-Oriented Locality Sensitive Hashing , *ACM International Conference on Multimedia (ACM MM),* Florence, Italy, 2010.

[68] K. Gao, Y. D. Zhang, W. Zhang, S. X. Lin, Affine Stable Characteristic Based Sample Expansion for Object Detection, *ACM International Conference on Image and Video Retrieval*, Xi'an, China, pp.422-429, 2010.

作者简介：

曹　娟：　中科院计算所，副研究员 caojuan@ict.ac.cn

张勇东：　中科院计算所，研究员

李锦涛：　中科院计算所，研究员，前瞻研究实验室主任